



ORIENTATION TO TEXT AND DATA MINING

JOHN SOUTHALL

BODLEIAN DATA LIBRARIAN & ECONOMICS SUBJECT CONSULTANT

MARCH 2025

AIMS OF TDM PRESENTATION

- Highlight that Text Mining and Data Mining tools are available
 - Has anyone been doing any data mining?
 - How many have worked with an API key?
 - How familiar are these terms to the group today?
- Discuss Bodleian subscribed platforms that support mining
 - Gale Digital Scholar Lab
 - TDM Studio
 - HathiTrust Digital Library

MINING CONTENT

- TDM approaches 'mine' content through;
 - Building up large volumes of text or data
 - Manipulating these corpora, databases or datasets
 - Using indirect, automated machine analysis - beyond the abilities of an individual or group
 - Revealing patterns, clusters, themes
 - Identifying new relationships in the content

ARRANGING AN API

- Use of an API (Application Programming Interface) is one way to compile a large dataset
- At its most **general** see an API as a gateway, contract or arrangement
 - Usually organized by publishers on an individual, case by case basis
 - Gives access to a dataset or text corpus beyond the usual licensing arrangement
 - Requires consideration of data security
 - Often needs input from an academic library
- Most commonly used for text mining of journal or newspaper content
 - Many suppliers do not allow mass downloads or data mining
 - Permission HAS to be sought in advance

SUPPORTING API USE

- An API is popular with researchers planning to use Python or other coding techniques
- Popular code libraries then support the analysis
 - NLTK
 - SpaCy (Python)
 - Tidytext (R)
- What is library role?
 - Making you aware of API availability from T&F, Sage etc.
 - Providing library approval or technical assistance

TYPICAL TDM TECHNIQUES

- Cluster Analysis
- Anomaly Detection
- Visualizations
 - Geographic
 - Topic Modelling
 - Sentiment Modelling
 - Stylometry: <https://en.wikipedia.org/wiki/Stylometry>
 - <https://computationalstylistics.github.io>

SUPPORTING TDM TECHNIQUES

- Library role is NOT to teach these, but..
 - Point to resources that inform them
 - Consider how library collections are being used (impact on usage figures)
- We can promote code or non-code TDM approaches
 - Where coding proficient researcher want to find sources
 - Where non-coding researchers want platforms that include data, API and codeless tools
- So the Bodleian is actively looking for platforms we can provide to you
 - Want feedback on your experience from the three in place so far

GALE DIGITAL SCHOLAR LAB

What Is the Gale Digital Scholar Lab?



The Lab is a single research platform where you can apply natural language processing tools to raw text data (OCR) from your institution's Gale Primary Sources holdings, or from uploaded OCR. Gale Digital Scholar Lab is organized in three broad steps: **Build, Clean, and Analyze**. These steps support newcomers and experienced users alike as they interpret both Gale Primary Sources and their own documents. An integrated Learning Center provides instructional tutorial videos and explanations throughout. The six built-in analysis tools are: Ngrams, Sentiment Analysis, Topic Modeling, Named Entity Recognition, Document Clustering, Parts of Speech.

CONTENT

Gale Primary Sources content

Provides access to millions of pages of content spanning many centuries and geographic regions, including:

- 17th and 18th Century Burney Collection
- 19th Century UK Periodicals
- Any Archives Unbound collections held by Oxford
- British Library Newspapers
- The Economist Historical Archive, 1843-2011
- Eighteenth Century Collections Online (ECCO)
- The Financial Times Historical Archive, 1888-2010
- The Illustrated London News Historical Archive, 1842-2003
- The Making of Modern Law: Legal Treatises, 1800-1926
- The Making of the Modern World 1450-1850
- Nineteenth Century U.S. Newspapers
- The Times Digital Archive
- Times Literary Supplement Historical Archive
- U.S. Declassified Documents Online.

GALE PRIMARY SOURCES

- Cross-search all available Gale digital archives from a single interface
- Explore history through digitized, text-searchable primary sources
- Discover new connections using Topic Finder and Term Frequency

Search across all selected Gale Primary Sources databases (33 of 33)

Start your research

Publication date(s)
 All Dates Before On After Between
 Include documents with no known publication date.

[View all limiters in Advanced Search >>](#)

Browse the Gale Primary Sources databases provided by your institution

Select All Deselect All

- GALE PRIMARY SOURCES
ARCHIVES OF SEXUALITY AND GENDER
Archives of Sexuality and Gender
Research a robust and significant collection of primary sources for the historical study of sex, sexuality, and gender.
- GALE PRIMARY SOURCES
BRITISH LITERARY MANUSCRIPTS ONLINE
British Literary Manuscripts Online
Explore a digitized collection of manuscripts of British authors that includes poems, plays, novels, diaries, and more.
- GALE PRIMARY SOURCES
DAILY MAIL HISTORICAL ARCHIVE, 1896-2004
Daily Mail Historical Archive
Search 100+ years of this major British newspaper, including access to the Atlantic Editions.

THE LAB

Library Menu: Bodleian Libraries of the University of Oxford

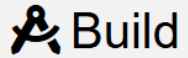
English

GALE DIGITAL SCHOLAR LAB

Build Clean Analyze My Research

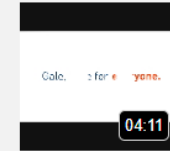


Add Note Search History



Build

Build Content Sets by finding documents in your institution's Gale Primary Sources holdings or by uploading your own documents.



Video: Building a Content Set

Additional topics:

- [Search Strategies](#)
- [Understanding Search Results](#)
- [View All Build Help »](#)

Search

- peter kemp
- peter kenyon, OCR confidence, and many more
- peter kelly
- peter kennedy
- peter ker
- peter kerr
- peter kendall
- peter kellner
- peter keenan
- peter kern
- peter kell

Upload

Drag and Drop .txt or .csv Files here or [Browse](#)

Get Upload Template

Create a Text Document

Manage All Uploads

BUILDING A CONTENT SET

Library Menu: Bodleian Libraries of the University of Oxford English

GALE DIGITAL SCHOLAR LAB Build Clean Analyze My Research

Search: peter kemp Add Note Add To Content Set Remove From Content Set Search History

7,054 Results **Sort by:** Relevance

Search Terms: Basic Search: peter kemp [Revise Search](#)

Select All

<p><input type="checkbox"/> Peter Kemp reflects</p> <p>OCR Confidence: Not Captured</p> <p>?Peter Kemp reflects on Edmund White's short stories, page 11 ?Peter Kemp reflects on Edmund White's short stories, page 11...</p>	<p>Publication Publication Date Pages</p> <p>The Sunday Times. March 12, 1995 1</p> <p>Archive Source library Content Type Document Type</p> <p>The Sunday Times Historical Archive Times Newspapers Limited Newspaper Front matter</p>
<p><input type="checkbox"/> Peter Kemp</p> <p>OCR Confidence: Not Captured</p> <p>PETER KEMP PETER KEMP Peter Kemp, DSO, MC, author, soldier and war reporter, died on October 30 aged 55. He was born in Bombay on August 19, 1915. PETER KEMP was in many ways a throwback to an earlier era. He...</p>	<p>Publication Publication Date Pages</p> <p>The Times. November 6, 1993 1</p> <p>Archive Source library Content Type Document Type</p> <p>The Times Digital Archive Times Newspapers Limited Newspaper Obituary</p>
<p><input type="checkbox"/> M'Lean Beats Peter Kemp</p> <p>Author: [By Cable to the Herald] OCR Confidence: 51%</p> <p>M LEAN BEATS PETER KEMP[BY CABLE TO THE HERALD.] M LEAN BEATS PETER KEMP[BY CABLE TO THE HERALD.] [] Sydney, Dec. 15.—A sculling match between Peter Kemp and McLean over the Parramatta course, for £300 a side and...</p>	<p>Publication Publication Date Pages</p> <p>The New York Herald (European Edition). December 16, 1890 1</p> <p>Archive Source library Content Type Document Type</p> <p>International Herald Tribune Historical Archive, 1887-2013 The New York Times Company Newspaper Article</p>

Filter Your Results

Archives	Content Type	Document type
Publication Date	Publication title	Publication Sections
Publication country or territory	Publication state/province	Publication city
Languages	Subjects	Author - items by
Source library	Illustrated works	OCR confidence range
Search Within		

Show Documents Added to

Select Content Set ▼

[Manage](#) [Analyze](#)

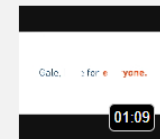
CLEANING



Clean

Select Cleaning Configuration to Edit

Default Cleaning Configuration



Video: Cleaning a Content Set
Additional topics:

- [Creating a Clean Configuration](#)
- [Applying During Analysis](#)
- [View All Clean Help »](#)

Stop words

Set the words you want the Analysis Tools to ignore. [Choose a Starter List](#)

a
about
above
across
after
afterwards
again
against
all
almost
alone
along
already
also
although
always
am
among

Text Correction

Options for automatic text correction that will be applied before each Analysis

- Turn on all options
- Text Modification**
 - All lower case
- Characters**
 - Remove all extended ASCII characters
 - Remove all number characters
- Special Characters**
 - Remove all special characters
 - [Set specific characters >](#)
- Punctuation**
 - Remove all punctuation
 - [Set specific punctuation >](#)

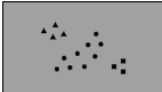
Configuration Notes

Space to make notes or describe the purpose of this configuration.

Configuration description/notes

ANALYSIS


Document Clustering 🔗


Document Similarity

Run Details

Default Setup ▼ Edit Run



Named Entity Recognition 🔗


Entities Found

Run Details

Default Setup ▼ Edit Run



Ngrams 🔗

 
Word Cloud Term Frequency

Run Details

(Mon Feb 24 10:12:50 UTC 2025) ▼ View New Setup Run Time: 1:06




Parts of Speech 🔗

 
Parts Comparison Pie Chart

Run Details

Default Setup ▼ Edit Run



Sentiment Analysis 🔗

  
Sentiment Scores Sentiment Over Time Sentiment By Timeframe

Run Details

(Mon Feb 24 10:03:34 UTC 2025) ▼ View New Setup Run Time: 0:27

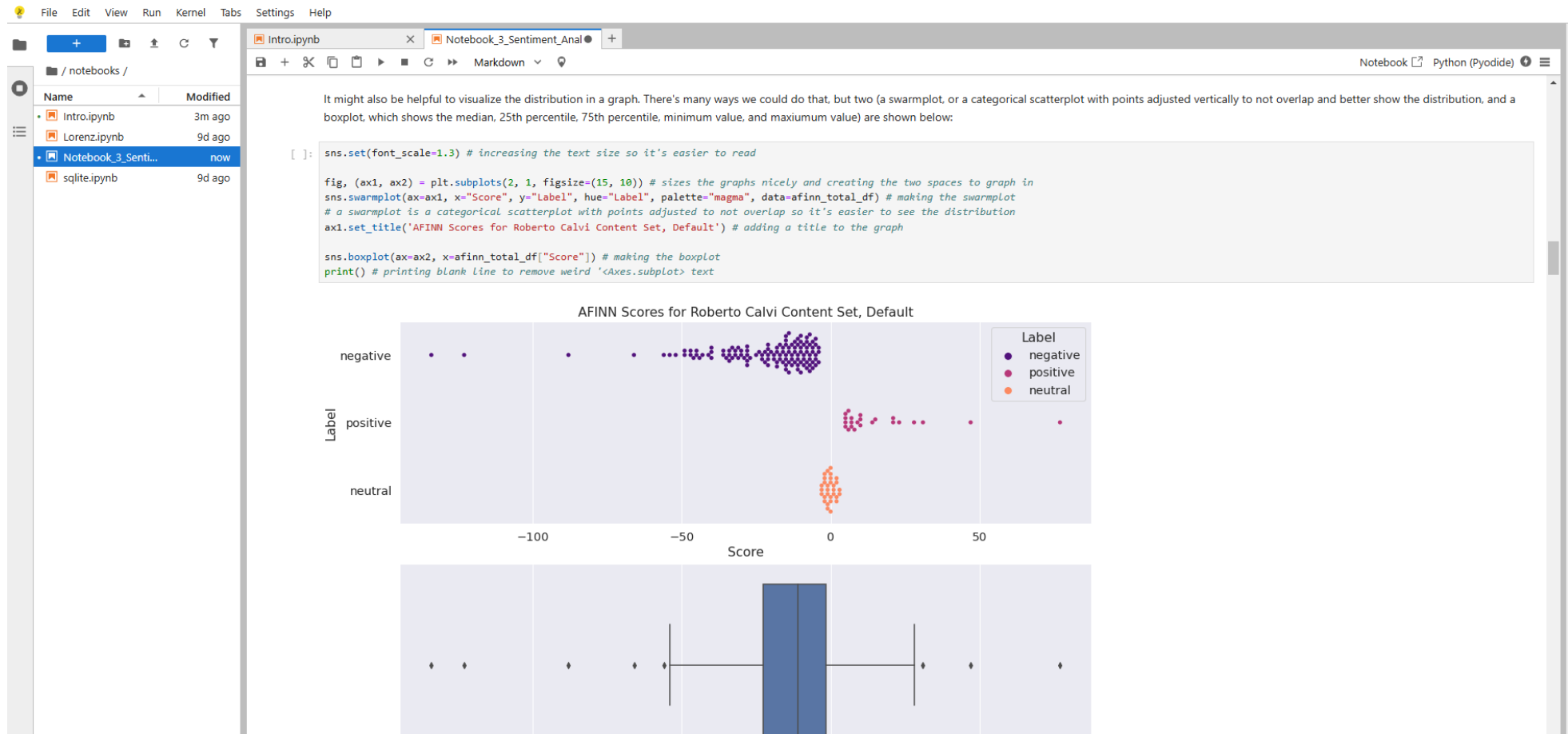
Topic Modeling 🔗

 
Topic Proportion Topics

Run Details

Default Setup ▼ Edit Run

RUNNING SCRIPTS IN BROWSER-BASED JUPYTER NOTEBOOK



GETTING STARTED

- Follow the instructions in the [Description on SOLO](#) or Databases A-Z:
- Click on *Link to Database*
- Click on *Institution Credentials...Recommended*
- Proceed with SSO
- Click on *Log In / Create Account*
- Click on *Use University Credentials* (preferred – Google log in also available)
- Select *Personal Workspace*

GALE DIGITAL SCHOLAR LAB - RESOURCES

- [Quick Start Guide](#) (video)
- [Learning Center](#) (documentation & videos – need to be logged into your workspace)
- [Support](#) (LibGuide, webinars & tutorials)



TDM STUDIO

Introduction

TDM Studio is a text and data mining tool created by ProQuest.

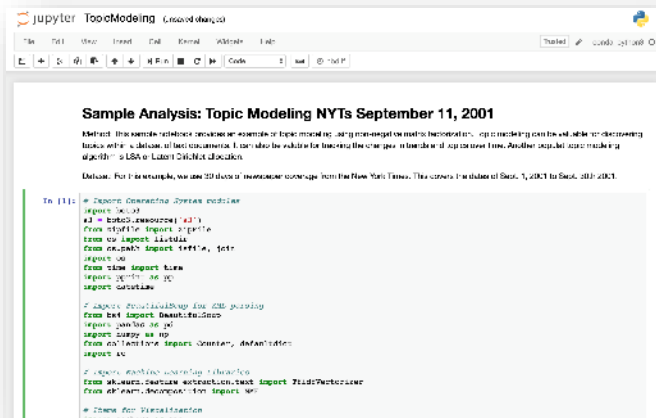
It allows programmatic analysis of published content from the millions of pages of news and scholarly publications provided through **current** university ProQuest subscriptions.



DATA ANALYSIS

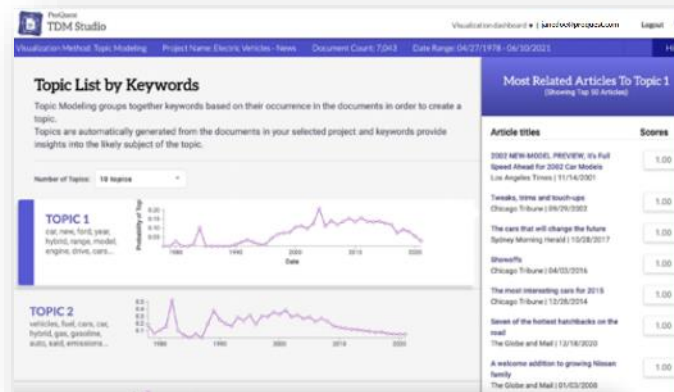
- TDM Studio provides three options for data analysis:

Workbench



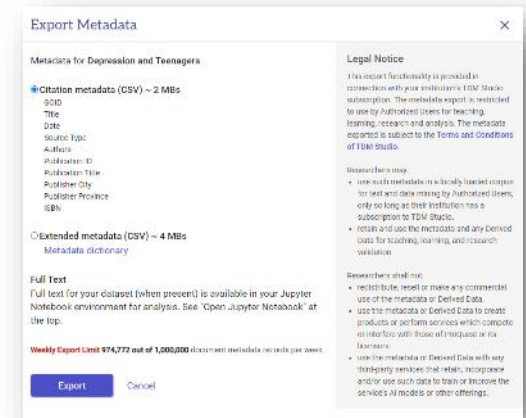
Cloud-based environment with Python and R libraries, preloaded scripts, and learning resources.

Visualization Dashboard



Analyze without coding using built-in geographic, topic, and sentiment analysis tools.

Metadata Export



No coding needed—export up to 1 million metadata records per week to use in preferred external tool such as Excel.

CONTENT

Content types:

- Dissertations – 6 million records
- Current and historical newspapers
- Congressional hearings
- Scholarly journals
- Historic periodicals
- Primary sources
- History Vault

Subjects:

- Art
- Business
- Health & Medicine
- History
- Literature
- Science & Technology
- Social Sciences

180+ Databases:

- U.S. Newsstream
- ProQuest Central
- ABI Inform Collection
- The Wall Street Journal
- Annual Register
- History Vault
- Nursing & Allied Health



DATASETS

Search across TDM Studio content and via the dataset creation page.

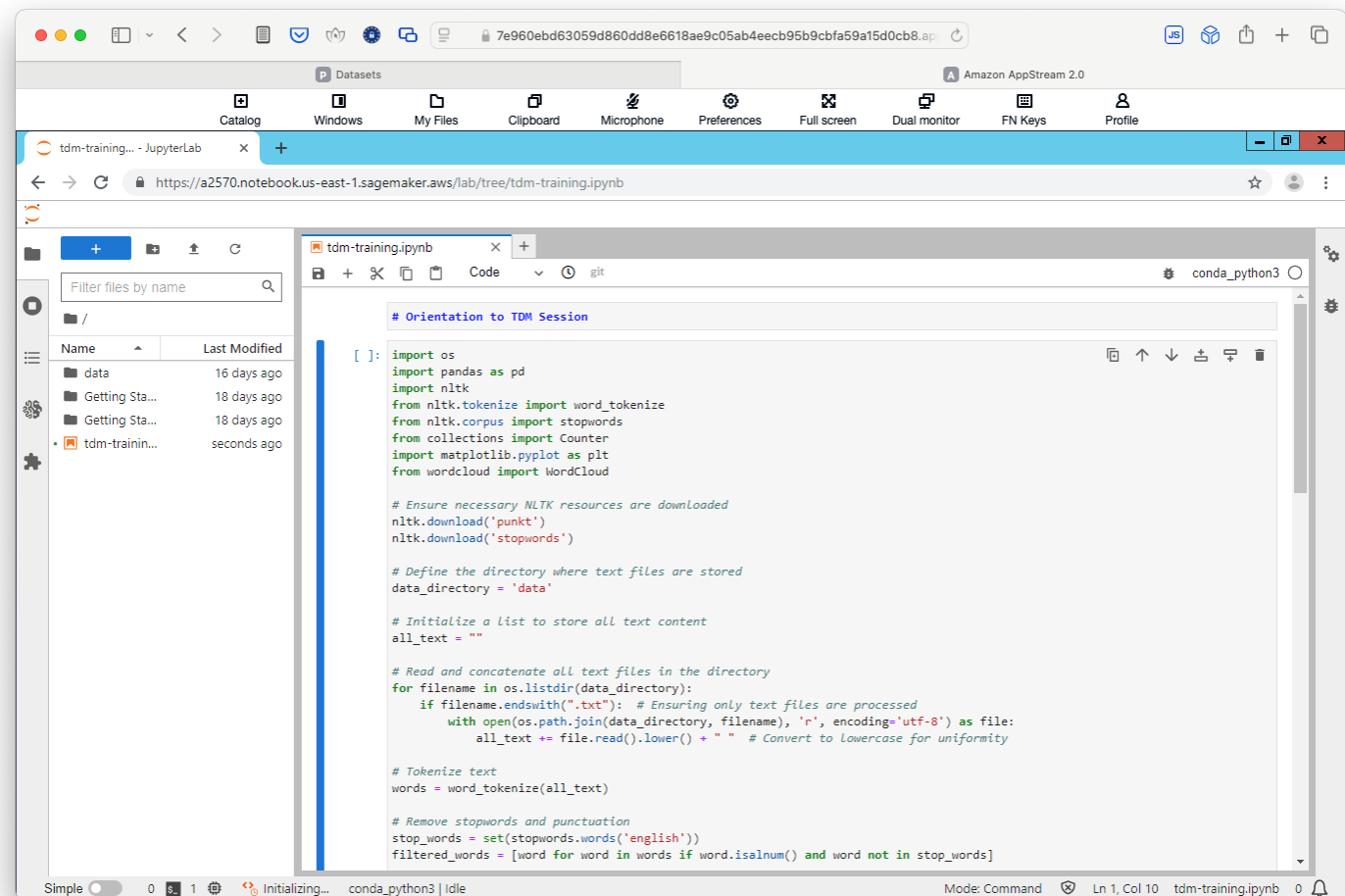
Benefits:

- **Search All Content:** Search across all available TDM Studio content – each dataset can be up to 2 million documents.
- **Easy Filtering:** Apply filters like publication and date on one page without restarting.
- **Quick Access to Leading Publications:** Instantly find popular publications for faster results.

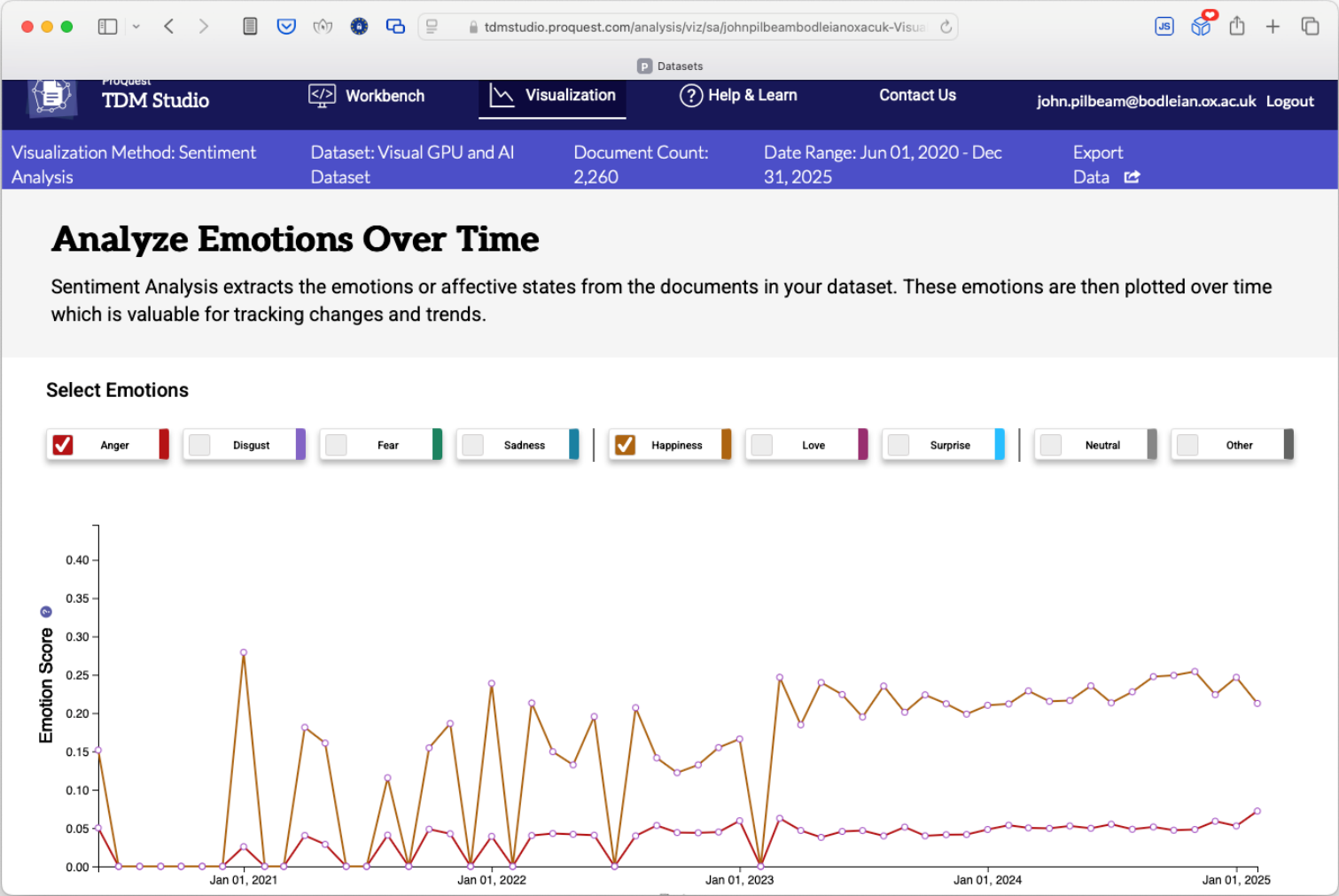
The screenshot displays the TDM Studio dataset creation interface. The search query is "AI AND GPU AND LLM". The page shows 854 documents with a "Save Dataset" button (Max 2 min documents). The left sidebar contains filters for "Applied filters" (ABI/INFORM Global, ABI/INFORM Trade & Industry, 1 Year), "Publication date" (1 Year, 5 Years, 10 Years, 50 Years), "Publication title" (View All), "Databases" (View All), and "Source Type" (Wire Feeds: 528, Trade Journals: 112). The main content area lists documents such as "Global Semiconductor 20 Feb 25_EN", "Nebius Group N.V. announces fourth quarter and full-year 2024 financial results", "Japan Telecommunications 19 Feb 25_EN", "Japan Electronic 19 Feb 25_EN", "'Where are the aliens?': Elon Musk explains Grok 3 mission, says xAI wants to answer the biggest questions [International-News]", "ETtech Explainer: Meet AI 'Saba' trained on Arabic, Tamil, Malayalam [Startups]", "Valentine's Day: Wall Street & the AI Bull Market on the Rocks?", and "WEEKLY RECAP: DATACENTERDYNAMICS LTD. NEWS THIS PAST WEEK FEB 13, 2025".

WORKBENCH

- Cloud-based environment
- Python and R libraries
(including PyTorch and TensorFlow)
- Preloaded scripts
- Learning resources



VISUALIZATION



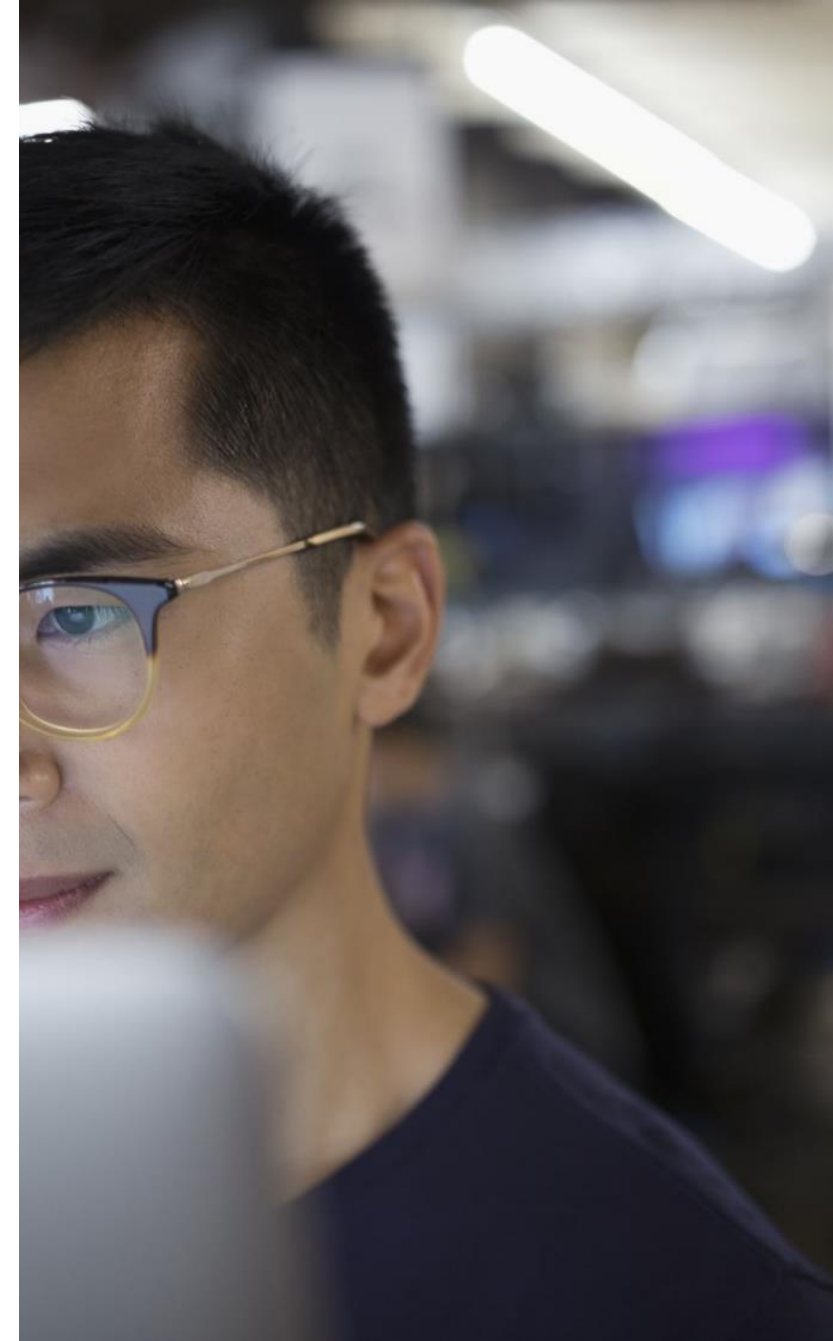
GETTING STARTED

Follow the instructions in the [Description on SOLO](#):

- Click on *Link to Database*
- Click on *Create Account*
- Register using your *University of Oxford* email address
- Check your email to verify and then log in to TDM Studio
- Log in to TDM Studio with your new account details at: <https://tdmstudio.proquest.com/home>

TDM STUDIO RESOURCES

- [Quick Start Guide](#)
- [TDM Studio Videos](#)
- [Documentation \(LibGuide\)](#)



HATHI TRUST DIGITAL LIBRARY

- The HT was founded in 2008 and acts as a digital library of fiction and non fiction
- Largest collections - Language & Literature, Philosophy, Religion, History, and Social Sciences
- English language titles make up 51% of the collection followed by German, French, Spanish, and Russian. The 400+ other languages include Chinese, Arabic, Japanese, and Afrikaans.
- 95% of the collection was digitized from print by Google and contributed by member libraries

HATHI TRUST DIGITAL LIBRARY

- Hathi Trust provides some titles that Google does not, such as digital collections and titles unique member institutions, and works from institutional repositories.
- Out of copyright content can be easily downloaded individually for close reading
- However TDM i.e. working with thousands of pages requires a special form of access through its research centre (HTRC)

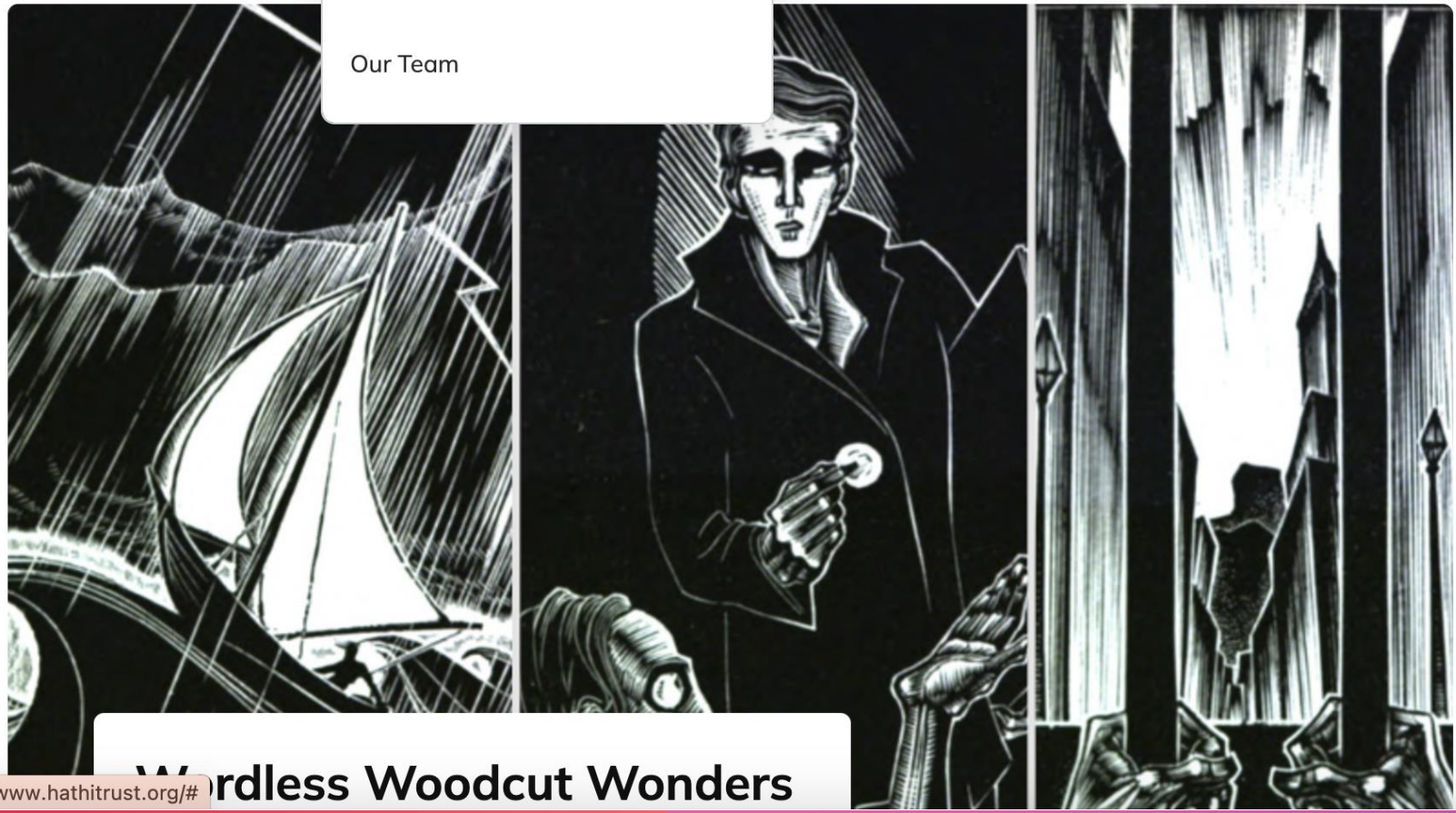
- Welcome to HathiTrust
- Our Mission & History
- HathiTrust Research Center (HTRC)
- Governance
- Our Team

Collection [Full Text & All Fields](#) [SEARCH](#)

You're searching in Full Text & All Fields

[Search Help](#) [Advanced Search](#)

FROM AROUND THE WORLD



<https://www.hathitrust.org/#wordless-woodcut-wonders>

A Farewell to Arms
by Ernest Hemingway, 1899-1961

Public Domain 2025



Vidas Cruzadas
by Jacinto Benavente, 1866-1954

Public Domain 2025



Hitty, Her First Hundred Years
by Rachel Field, 1894-1942

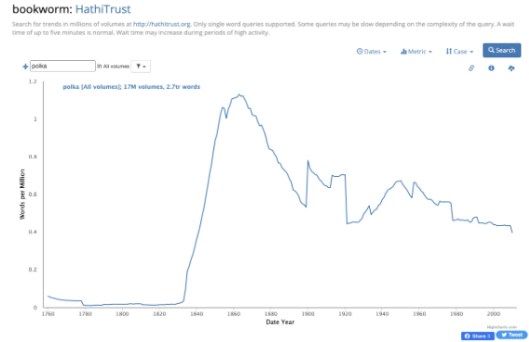
Public Domain 2025



Text and data mining tools

HTRC has developed a suite of tools and services for text data mining including web-based algorithms, freely-accessible datasets, and secure computing capsules. Access the tools below, as well as tutorials and other documentation on [HTRC Analytics](#).

Text and Data Mining Tools



HathiTrust + Bookworm

Visualizes word trends in millions of volumes held by HathiTrust. It enables scholars to discover new textual use patterns across the entire corpus, including in-copyright and public domain volumes.

Algorithms

HTRC Algorithms are click-to-run tools for text analysis. They require no programming, and researchers can set the parameters for their analysis. Use them to explore HathiTrust worksets, which are groups of titles from the collection.

Datasets and Data APIs

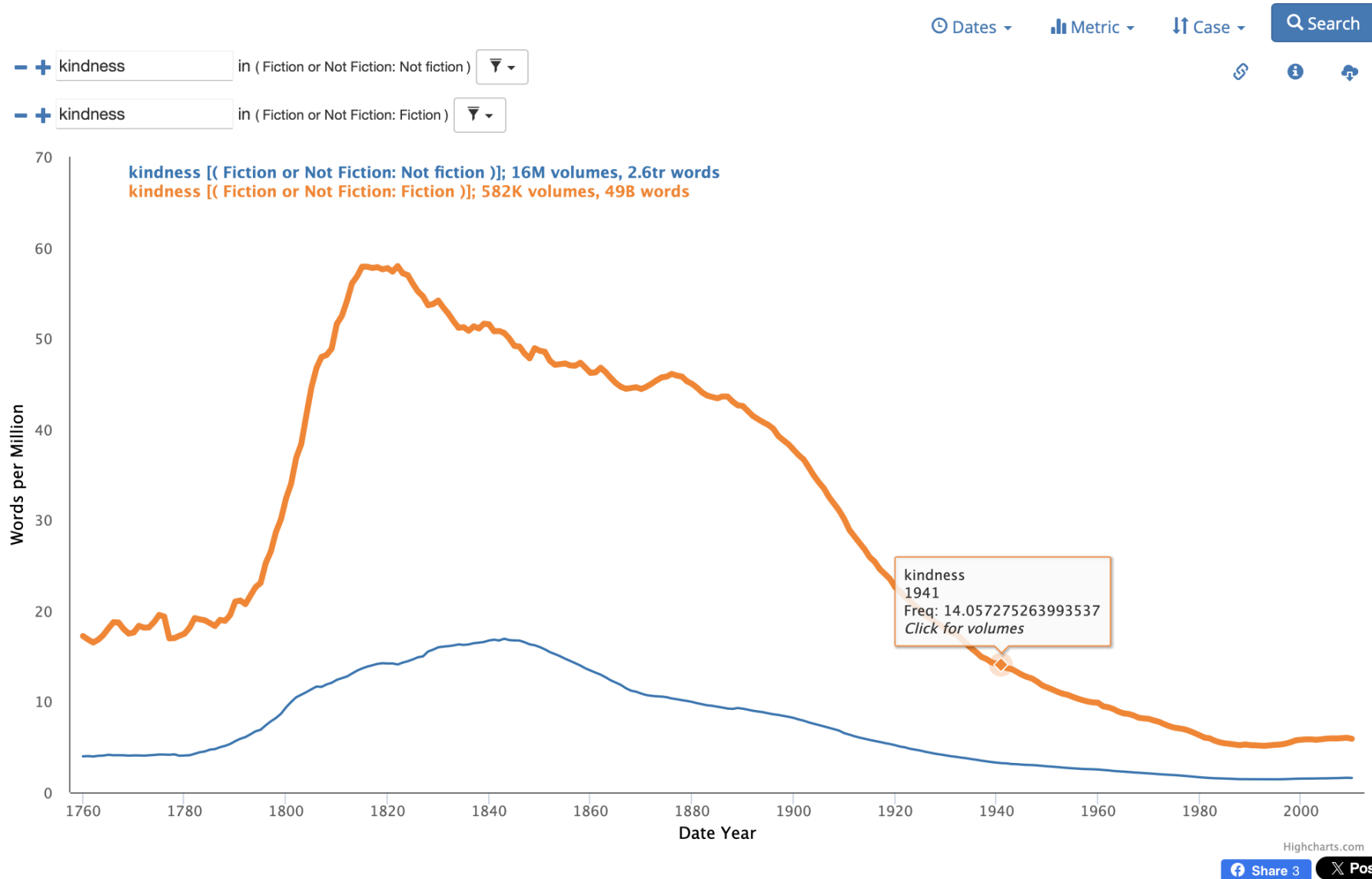
A dataset compiles specific features extracted from the full text of the HathiTrust corpus. Certain datasets are freely available and pre-created. HathiTrust also provides access to several data APIs. Other data and datasets can be accessed upon request.

Data Capsules

An advanced computing environment available only to HathiTrust members, data capsules provide high-capacity computing for advanced text an. Access to in-copyright material is available to HathiTrust members.

bookworm: HathiTrust

Search for trends in millions of volumes at <http://hathitrust.org>. Only single word queries supported. Some queries may be slow depending on the complexity of the query. A wait time of up to five minutes is normal. Wait time may increase during periods of high activity.



TOOLS AND SKILL LEVELS

Tool	Level	Data Format	Method
Bookworm	Beginner	None – results only	Visualisation
Web Algorithms	Beginner	None – results only	Topic Modelling, Named Entity Extraction etc
Extracted Features dataset	Intermediate	Word counts in structured file	Any corpus based
Data Capsule	Advanced	Raw text	Any corpus based

Voyant Tools x Home x header-footer-remover- x +

localhost:8889/notebooks/header-footer-remover-sample-notebook-exp.ipynb

jupyter header-footer-remover-sample-notebook-exp Last Checkpoint: Last Thursday at 11:33 PM (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

```

In [1]: import re
import os
import glob
import shutil
from collections import defaultdict
from typing import List, TypeVar, Set, Iterator, Optional, Tuple, Dict

from htrc.models import Page, PageStructure, HtrcPage
from htrc.utils import clean_text, levenshtein, pairwise_combine_within_distance, flatten
from htrc.runningheaders import parse_page_structure

In [2]: def load_vol(path: str, num_pages: int) -> List[HtrcPage]:
pages = []
py_num_pages = num_pages-1
for n in range(py_num_pages):
    if n == 0:
        n = 1
        page_num = str(n).zfill(8)
        with open('{} / {}'.format(path, page_num), encoding='utf-8') as f:
            lines = [line.rstrip() for line in f.readlines()]
            pages.append(HtrcPage(lines))
    else:
        page_num = str(n).zfill(8)
        with open('{} / {}'.format(path, page_num), encoding='utf-8') as f:
            lines = [line.rstrip() for line in f.readlines()]
            pages.append(HtrcPage(lines))

return pages

In [3]: # UPDATE DIRECTORY: REPLACE 'EXPANDED' WITH THE NAME OF YOUR WORKSET FOLDER
vol_path_list = glob.glob('TPAINECommon/*')

vol_path_list

Out[3]: ['ChrisGadDocs/uc1.b4349532']

In [4]: for vol_path in vol_path_list:

```

```

Command: /usr/lib/jvm/java-8-oracle
-Djava.io.tmpdir=/opt/applications
lications/VoyantServer2_4-M7/VoyantS
ime 8888 /opt/applications/VoyantSe

```

TPAINEComm

secure_volume TPAINECo

Recent

Home

Desktop

Documents

Downloads

Music

Pictures

Videos

Trash

Network

Computer

Floppy Disk

release_spool

secure_volume

Connect to Server

Name

msu.312

Cleaning data in Jupyter

Open data in Voyant

Terminal window showing command: `/usr/lib/jvm/java-8-oracle...`

File manager window showing `secure_volume` with folders like `Loyalist`, `HLauren`, `SAdams`, `htrc`, `mylist.t`, `lost+fo`, `data`, `header-`, `msu.312`, `volume-`, `TPaineC`, `TPaineC`, `ChrisGa`

Voyant Tools interface showing a corpus analysis of a document. The browser address bar shows `127.0.0.1:8888/?corpus=7806db3c6fa953d83b0d1a524d657d15`.

Navigation tabs: Cirrus, Terms, Links, Reader, TermsBerry, Trends, Document Terms.

Term	Count	Trend
land*	14	
manufacture*	14	
mechanic*	14	
planter*	14	
merchant*	14	

Text snippet: "ing Charleston in 1780, wrote a pamphlet, The Candid Retrospect of the American War Examined by Whig Principles: DAB, XVII, 357-8; Carl Becker, History of Political Parties in the Province of New York (Madison : University of Wisconsin, 1909), pp. 38-39; Morgans, op. cit., p. 184; William H. W. Sabine, Historical Memoirs of William Smith ... (New York: np, 1956), pp. 29-36. 69 without stamps. In this they overruled the Chief Justice," whose character and abilities (if he has either), you cannot be unacquainted with, who was of a different opinion. But the clerk of the Common Pleas, Mr. Dougal Cambell, refused to do his part, and make an entry they ordered; they appointed another to do it, and out of tenderness to him, did not commit him for his refractoriness, not being aware of such a refusal and, expecting he would think better of it the next day, or if he did not, that the Lieut. Gov. upon their application, would suspend him, accordingly the next day finding

Summary: This corpus has 1 document with 20,091 total words and 4,067 unique word forms. Created now. Vocabulary Density: 0.202. Average Words Per Sentence: 28.4. Most frequent words in the corpus: gadsden (66); carolina (52); committee (50); congress (45); william (45)

Contexts table:

Document	Left	Term	Right
1) Chris...	settlers of '4 upon the occaul constl...	im...	, and upon YOU the taxes and burth...
1) Chris...	a wise and prudent body of men will...	im...	; or if they do, what must be the co...
1) Chris...	dispatch are absolutely necessary. ...	im...	all that you can, and think it probabl...
1) Chris...	had drank deeped . I am per- We ar...	im...	to Great Britain as any upon the co...
1) Chris...	the Common Cause and as it were i...	im...	of a union amongst all the colonies ...

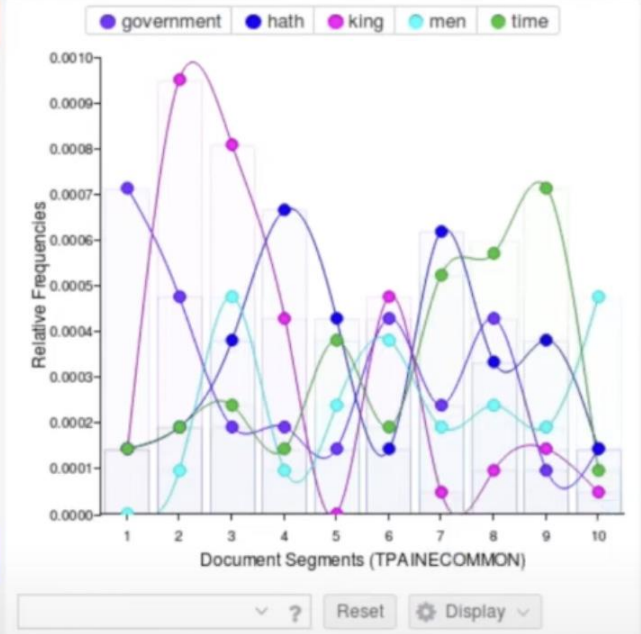
Line graph showing Relative Frequencies (0.0000 to 0.0012) across Document Segments (1 to 10) for terms: carolina, committee, congress, gadsden, william.

Terms filter: manufacture*, mechanic*, merchant*, import*, land*



TPAINECOMMON

E
211
.P1455
1995
COMMON
SENSE
THOMAS
PAINE
GREAT BOOKS IN PHILOSOPHY
E
211
.P1455
1995
COMMON
SENSE
THOMAS



Terms:

This corpus has 1 document with 21,037 total words and 3,899 unique word forms. Created now.

Vocabulary Density: 0.185

Average Words Per Sentence: 33.6

Most frequent words in the corpus: **hath** (72); **time** (67); **king** (66); **government** (64); **men** (50)

Document	Left	Term	Right
1) TPAI...	as the king of England	hath	undertaken in his own right
1) TPAI...	the following sheets, the author	hath	studiously avoided every thing which
1) TPAI...	every man to whom nature	hath	given the power of feeling
1) TPAI...	of the people; but this	hath	all the distinctions of a
1) TPAI...	fate of Charles the First	hath	only made kings more subtle
1) TPAI...	Holland, with- out a king,	hath	enjoyed more peace for the
1) TPAI...	kings, and the Christian world	hath	improved on the plan by
1) TPAI...	equivocal construction. That the Al...	hath	here entered his protest against

Voyant Server

File Help

Stop Sen
Open W

Voyant Server

Console updated: Tue, 2 Feb 2021 19:...

Please provide the jar on your classpath to...
See tika-parsers/pom.xml for the correct ve...
log4j:WARN No appenders could be found for...
log4j:WARN Please initialize the log4j system...
log4j:WARN See http://logging.apache.org/lo...

2021-02-02 19:56:37.716:INFO/: qtp9891100
2021-02-02 19:56:37.836:INFO/: qtp9891100
2021-02-02 19:56:37.934:INFO/: qtp9891100
2021-02-02 19:56:39.287:INFO/: qtp9891100
2021-02-02 19:56:39.321:INFO/: qtp9891100
2021-02-02 19:56:39.344:INFO/: qtp9891100
2021-02-02 19:56:39.460:INFO/: qtp9891100
2021-02-02 19:56:39.539:INFO/: qtp9891100
2021-02-02 19:56:39.577:INFO/: qtp9891100
2021-02-02 19:56:39.748:INFO/: qtp9891100

Downloads
Music
Pictures
Videos
Trash
Network
Computer
Floppy Disk
release_pool
secure_volume
Connect to Server

HLaurer
SAdams
htcr
mylist.b
lost+fo
data
header-
TPAINE
volume-
TPaineC
TPaineC
ChrisGa

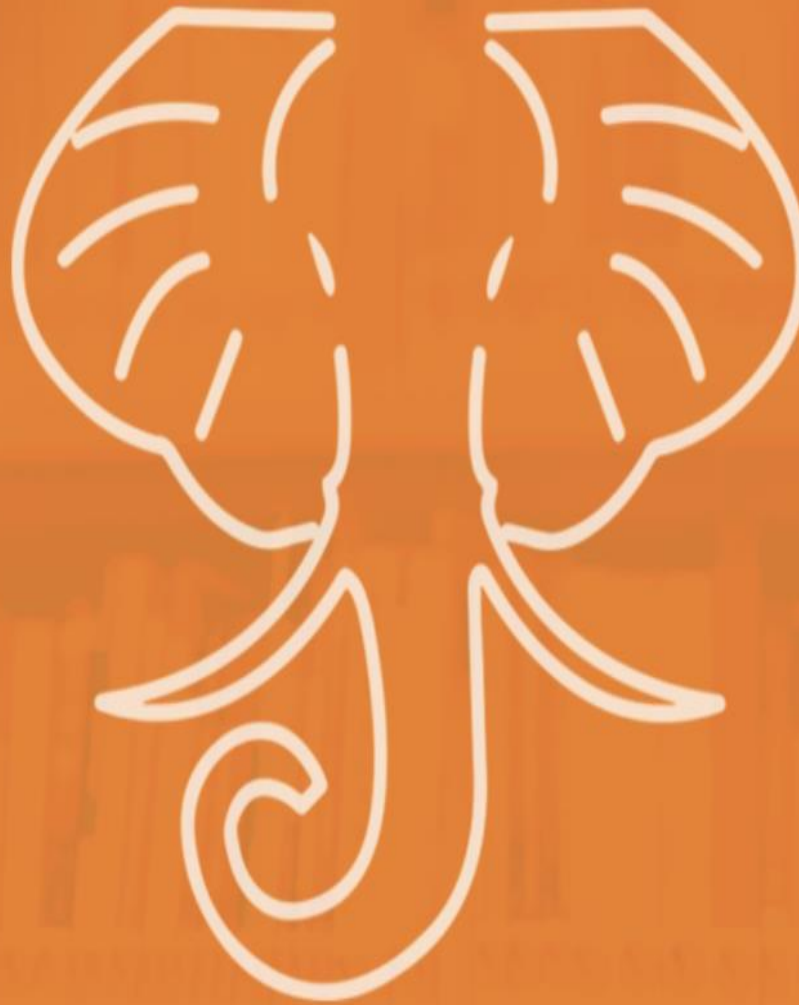
GETTING STARTED

Follow the instructions in the [Description on SOLO](#):

- Click on *Link to Database*
- Click on *Log In*
- Select *University of Oxford* as institution
- Select HTRC from 'About' list
- Click on HTRC services link and follow prompts

HTRC RESOURCES

- [Quick Start Guide](#) (Text as data)
- [Tom Paine Analysis Video](#)
- [Documentation \(LibGuide\)](#) (HTRC Analytics)



SUPPORTING TDM AT OXFORD

- We see TDM as another facet of information skills competence
 - Using content in new ways
 - Overlapping with Research Data Management
 - Requiring responsible use of data
 - Awareness of licensing and restrictions on use
- Make use of the recently published TDM Libguide
- Currently accessed via [Bodleian Data Service](#)



SOLO

Books, journals, databases

Reading lists ▶

Exam Paper Archive

Subject and research guides

Special collections ▶

Bodleian Data Service ▼

About the Bodleian Data Service

Finding data and statistics ▶

Accessing and using data ▶

Text and data mining

Training and support ▶

Bodleian Data Service

The Bodleian Data Service provides a range of services for researchers and students at the University of Oxford who need to make secondary use of statistics and data.

For the academic year 24/25 there is a new data access service supporting researchers undertaking Text and Data Mining (TDM), with the aim of removing barriers and providing an easy-to-access pathway.

- Finding data and statistics ▶
- Accessing and using data ▶
- Text and data mining ▶
- Training and support ▶
- About the Bodleian Data Service ▶



Text and Data Mining pilot project: TDM training

Social Sciences Division and Humanities

[Home](#)[Current TDM-ready collections](#)[TDM training](#)[TDM software](#) 

UKDS TDM training

UK Data Service Webinars

Text-mining is one of many data-mining techniques that social scientists are using to turn unstructured (or more accurately, semi-unstructured) material into structured material that can be analysed statistically. In this way, researchers are gaining access to new materials and methods that were previously unavailable. As such, it is increasingly important that social scientists have a clear understanding of what text-mining is (and what is isn't) as well as how to use text-mining to achieve some basic and more advanced research outcomes.

- [Introduction to Text-Mining](#)

This webinar is the first in a series of three on understanding and using text-mining methods within social science research contexts.

The first webinar covers the concepts behind fully structured and semi-unstructured data, the theory behind capturing and amplifying existing structure, and the four basic steps involved in any text-mining project.

- [Text-Mining: Basic Processes](#)

Webinar two will dive into the steps needed to do some of the most common text-mining analyses and will be accompanied by an online interactive notebook that allows participants to see, edit and execute the demonstrated code.

- [Text-Mining: Advanced Options](#)

Webinar three rounds off the series by diving into the concepts behind more advanced text-mining analyses, presenting some sample code that participants may find useful, and introducing some work that provides further learning opportunities. This webinar will also be accompanied by an online interactive notebook that allows participants to see, edit and execute the demonstrated code.

Text and Data Mining (IBM)

"**Text mining** is the practice of analyzing vast collections of textual materials to capture key concepts, trends and hidden relationships" - [What is text mining?](#)



"**Data mining** is the use of machine learning and statistical analysis to uncover patterns and other valuable information from large data sets" - [What is data mining?](#)



CLOSING

- Thank you
- Any questions?
- Feel free to get in touch for support in the future
 - john.southall@bodleian.ox.ac.uk
 - <https://www.bodleian.ox.ac.uk/collections-and-resources/data>